Adaptive pairwise comparison for educational measurement

Crompvoets, Elise; Béguin, A.A.; Sijtsma, K.

Published in:
Journal of Educational and Behavioral Statistics

DOI:
10.3102/1076998619890589

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–23
DOI: 10.3102/1076998619890589
Article reuse guidelines: sagepub.com/journals-permissions

Adaptive Pairwise Comparison for Educational Measurement

Elise A. V. Crompvoets
Tilburg University
Cito

Anton A. Béguin
Cito

Klaas Sijtsma
Tilburg University

Pairwise comparison is becoming increasingly popular as a holistic measurement method in education. Unfortunately, many comparisons are required for reliable measurement. To reduce the number of required comparisons, we developed an adaptive selection algorithm (ASA) that selects the most informative comparisons while taking the uncertainty of the object parameters into account. The results of the simulation study showed that, given the number of comparisons, the ASA resulted in smaller standard errors of object parameter estimates than a random selection algorithm that served as a benchmark. Rank order accuracy and reliability were similar for the two algorithms. Because the scale separation reliability (SSR) may overestimate the benchmark reliability when the ASA is used, caution is required when interpreting the SSR.

Keywords: *adaptive measurement; comparative judgment; holistic measurement; pairwise comparison*

Pairwise comparison is a method that allows measurement of an attribute by means of comparison of objects with respect to the attribute in pairs. Models for pairwise comparison data are used to obtain a scale for the objects with respect to the attribute. The method was first introduced by Thurstone (1927). Objects may be anything such as sports teams or product brands (see Cattelan, 2012, for an overview of applications outside education), but in educational measurement, objects are mostly students' responses to an assignment or an examination. The assignment or the examination is used to measure an attribute of the students, and the students' responses give an indication of their attribute level. For example, to create a rank order of students with respect to creative thinking skills, primary school teachers compare students' responses to a creative thinking assignment

with each other and rate which of two students showed the highest level of creative thinking. Because people perform pairwise comparisons routinely on a daily basis, for example, when deciding to eat a salad or a burger for lunch, pairwise comparison is highly intuitive and provides a natural task for people to perform. Laming (2004) even argued that every decision we make is based on comparative judgment. The advantage of using an everyday process in an assessment task is that people, including raters, are familiar with it, resulting in relatively fast and time-efficient judgment.

In educational measurement, pairwise comparison is becoming an increasingly popular assessment method (Bramley & Vitello, 2018; Lesterhuis, Verhavert, Coertjens, Donche, & De Mayer, 2017). The method has been used in a variety of contexts, ranging from art assignments (Newhouse, 2014) to academic writing (Van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2016) and mathematical problem-solving (Jones & Alcock, 2013). The examples we mentioned are by no means exhaustive (e.g., Bartholomew, Strimel, & Yoshikawa, 2018; Seery, Canthy, & Phelan, 2012; Steedle & Ferrara, 2016) but give an impression of the wide range of contexts where pairwise comparison has been used. These contexts have in common that the attribute of interest cannot easily be divided into smaller attribute aspects that validly cover the total attribute. For example, creativity of an art assignment is more than a summary of aspects of the art assignment, such as use of colors and shapes, and assessing such aspects separately would not add up to an assessment of creativity. For this reason, in these contexts, the attribute of interest is difficult to measure validly using conventional analytic measurement methods such as rubrics or criteria lists (Van Daal et al., 2016). Pairwise comparison is a promising approach for measuring these attributes because evaluation can take place in a holistic manner (i.e., evaluating attributes as a whole; Sadler, 2009). Some authors argue that pairwise comparison should replace conventional analytic assessment methods for all assessments (Pollitt, 2004, 2012) because the method can reduce costs in terms of time, money, or both and may even improve scales' measurement properties (Bramley & Vitello, 2018; Pollitt, 2012).

Unfortunately, the large number of pairwise comparisons required for reliable measurement counteracts the time-efficiency advantage of making each comparison in a short amount of time. We need a sufficient number of comparisons to estimate the probabilities that the objects are preferred to the other objects accurately. In addition, to avoid capitalization on sample results, the selected comparisons should be a representative sample of all possible comparisons. Although each comparison often takes little time, it is unfeasible to ask of raters or teachers to compare assignments of all students to all other students because a small class of 20 students already provides 190 unique comparisons. The discrepancy of the interests of reliable measurement and low rater burden creates an efficiency–reliability trade-off (Bramley & Vitello, 2018; Lesterhuis et al., 2017), and deciding on the number of comparisons to present to the raters is

an important issue with respect to this trade-off. For an elaborate discussion about labor costs and timings, see Steedle and Ferrara (2016).

Making the comparison process adaptive is the most prominent approach to influence the efficiency–reliability trade-off positively (Bramley & Vitello, 2018; Pollitt, 2012). Adaptive pairwise comparison entails that the objects that are presented to the rater are selected to provide optimal information about the rank order of the objects. Which objects are selected is determined based on the information obtained in previous comparisons. The approach has similarities with computerized adaptive testing (e.g., see Van der Linden & Glas, 2010; Wainer et al., 2000), in which each next item is selected based on the estimated ability of a test taker as measured using the items administered thus far. Using adaptive pair selection, the same reliability should be achieved using fewer comparisons than using the common random pair selection. The challenge is efficiently selecting object pairs to be compared while the estimates of the object parameters still have relatively large standard errors. Unfortunately, current algorithms, for example, the Swiss method and the adaptive method discussed in Pollitt (2012) and a combination of the two (Pollitt, 2015; Rangel-Smith & Lynch, 2018), do not sufficiently take the uncertainty of the object parameters into account. Consequently, the algorithms may inflate the scale separation reliability (SSR) coefficient (Bramley, 2015; Bramley & Vitello, 2018), which is the ratio of the estimated true variance of the object parameters and the observed variance of the object parameter estimates, thereby overestimating reliability. As a result, the reliability may be overestimated, but Rangel-Smith and Lynch (2018) claim that the SSR inflation is mitigated when their adaptation of the adaptive algorithm is used with a sufficiently large number of comparisons.

In this study, we developed an adaptive selection algorithm (ASA) that takes the uncertainty of the object parameters into account when selecting the next object pair. We conducted a simulation study to investigate the performance of the algorithm and compared it with the performance of a random selection algorithm. The performance of the selection algorithms was evaluated by means of the uncertainty of the object parameters, the rank order accuracy, and the reliability. In general, we expected that the ASA would perform better than the random selection algorithm on all three evaluation criteria: lower uncertainty of the object parameters, higher rank order accuracy, and higher reliability. We varied the number of objects to be compared and the proportion of the total number of unique comparisons.

This article is organized as follows. First, we discuss the ASA algorithm in more detail. Second, we describe the steps of the parameter estimation procedure. Third, we describe the simulation study, and fourth, we discuss the results. Fifth, we discuss some exploratory analyses, and we end with a discussion.

ASA

The goal of the ASA is to select in each step the most informative pair of objects for the rater to compare given the results of previous comparisons. More specifically, the object of which the parameter estimate has the largest standard error is selected, so the next comparison provides information about the parameter about which we are most uncertain. This object is compared with an object of which the parameter has a high probability to be close to the parameter of the selected object on the latent variable scale. This selection procedure not only provides most information, but it also creates a connected network of comparisons as quickly as possible. We are most uncertain about objects that were not compared before, they are closest to other objects that were not compared before (in the middle of the scale), and subsequently the groups of comparisons are linked via comparison of two previously preferred objects or two previously nonpreferred objects. The algorithm is constrained to let all unique comparisons occur only once to prevent undesirable dependencies that may arise between comparisons of the same pair of objects. This restriction corresponds formally with a single rater that performed all comparisons. We elaborate on this choice in the discussion.

The algorithm is an iterative process using the following steps. First, the object parameter estimates based on the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) were computed using the data collected up to this point. Let N be the number of objects, and let i and j ($i,j = 1, \ldots, N$) be object indices. The BTL model defines the probability that object i is preferred to object j in a paired comparison by means of

$$P(i > j|\theta_i, \theta_j) = \frac{\exp(\theta_i - \theta_j)}{1 + \exp(\theta_i - \theta_j)},$$

where θ_i and θ_j are the attribute parameters of objects i and j, respectively. Second, the standard errors corresponding to these estimates based on the observed Fisher information were computed. Third, object i was selected from the objects that had not previously been compared to all possible objects j and that had the largest standard error. Subsequently, object j was selected from the objects that had not previously been compared to object i. Selection probabilities for these objects were computed by deriving the probability densities at the object parameter locations of a normal distribution with estimated mean $\hat{\theta}_i$ and estimated standard error $SE\left(\hat{\theta}_i\right)$, so that $\theta_i \sim \mathcal{N}\left[\hat{\theta}_i,\ SE\left(\hat{\theta}_i\right)\right]$. Dividing the density of each possible object j by the sum of the densities of all possible objects j created probabilities. Using probabilities for selection rather than directly selecting pairs based on estimates of the distances between objects on the attribute scale, the algorithm takes the uncertainty of the object parameters into account. After the two selected objects were compared, the data and comparison

counts were updated, and the selection algorithm steps were repeated until a predefined stopping criterion was reached. It may be noted that the stopping criterion is not inherent to the algorithm and can be chosen differently in different situations (e.g., different numbers of comparisons as described in the Method section).

Parameter Estimation

The object parameters were estimated using maximum likelihood (ML). To be able to obtain parameter estimates of objects that are preferred in all comparisons or objects that are preferred in none of the comparisons, that is, objects with perfect scores or zero scores, respectively, 0.01 prior observation was added to each possible outcome of each possible comparison between two different objects. This small addition of data has an almost negligible impact on the parameter estimates, and the impact decreases even further when the number of performed comparisons increases.

Let n_i be the total number of comparisons including object i, x_i be the number of comparisons in which object i is preferred, x_{ij} be the number of comparisons in which object i is preferred to object j, X be the data matrix containing all x_{ij}, and $\boldsymbol{\theta}$ be the vector of object parameters. The likelihood of the BTL model, including all comparisons performed, can be written as

$$\begin{aligned} L(\boldsymbol{\theta}|X) &= \prod_{i}^{N}\prod_{\neq j}^{N}\left(\frac{e^{\theta_i-\theta_j}}{1+e^{\theta_i-\theta_j}}\right)^{x_{ij}} \\ &= \prod_{i}^{N}\prod_{\neq j}^{N}\frac{e^{x_{ij}(\theta_i-\theta_j)}}{(1+e^{\theta_i-\theta_j})^{x_{ij}}} \\ &= \frac{e^{\sum_{i=1}^{N}(2x_i-n_i)\theta_i}}{\prod^{N}\prod_{i\neq j}^{N}(1+e^{\theta_i-\theta_j})^{x_{ij}}}. \end{aligned} \tag{1}$$

It may be noted that the whole fraction is raised to the power x_{ij} because the product is taken across both i and j, and every comparison should occur in the equation once. The log likelihood can be obtained by taking the natural logarithm of the likelihood function, so that

$$\begin{aligned} \log L(\boldsymbol{\theta}|X) &= \log\left(\frac{e^{\sum_{i=1}^{N}(2x_i-n_i)\theta_i}}{\prod^{N}\prod_{i\neq j}^{N}(1+e^{\theta_i-\theta_j})^{x_{ij}}}\right) \\ &= \sum_{i=1}^{N}(2x_i-n_i)\theta_i - \sum_{i}^{N}\sum_{\neq j}^{N}x_{ij}\cdot\log\left(1+e^{\theta_i-\theta_j}\right). \end{aligned} \tag{2}$$

The log likelihood was optimized following a minorization–maximization algorithm that belongs to the subset of expectation–maximization algorithms

(Hunter, 2004). Let n_{ij} be the number of comparisons between object i and object j. For iteration $k = 1, 2, \ldots, K$, let

$$\theta_i^{(k+1)} = \log\left\{ x_i \cdot \left[\sum_{j\,:\,j \neq i}^{N} \frac{n_{ij}}{e^{\theta_i\,(k)} + e^{\theta_j\,(k)}} \right]^{-1} \right\}. \tag{3}$$

To identify the model, if the resulting vector $\boldsymbol{\theta}^{(k+1)}$ did not have a mean of 0, it was renormalized as

$$\theta_i^{(k+1)} = \theta_i^{(k+1)} - \frac{\sum_{i=1}^{N} \theta_i}{N}. \tag{4}$$

The standard errors corresponding to the ML estimates of the object parameters were computed as the inverse of the observed Fisher information $\mathcal{I}(\theta)$. To obtain $\mathcal{I}(\theta)$, we first derived the gradient of the log likelihood. For each θ_i, this was equal to

$$\frac{\partial \log L(\boldsymbol{\theta}|X)}{\partial \theta_i} = 2x_i - n_i - \sum_{j\,:\,j \neq i}^{N} \left(\frac{x_{ij} \cdot e^{\theta_i - \theta_j}}{1 + e^{\theta_i - \theta_j}} - \frac{x_{ji} \cdot e^{\theta_j - \theta_i}}{1 + e^{\theta_j - \theta_i}} \right). \tag{5}$$

Second, we derived the second partial derivative of the log likelihood to θ_i, resulting in

$$\frac{\partial^2 \log L(\boldsymbol{\theta}|X)}{\partial \theta_i^2} = -\sum_{j\,:\,j \neq i}^{N} \left(\frac{x_{ij} \cdot e^{\theta_i - \theta_j}}{\left(1 + e^{\theta_i - \theta_j}\right)^2} + \frac{x_{ji} \cdot e^{\theta_j - \theta_i}}{\left(1 + e^{\theta_j - \theta_i}\right)^2} \right). \tag{6}$$

The standard errors of the object parameter estimates can then be computed as

$$\begin{aligned} S^2\left(\hat{\theta}\right) &= \frac{1}{\sqrt{\mathcal{I}(\theta)}} \\ &= \frac{1}{\sqrt{-\dfrac{\partial^2 \log L(\boldsymbol{\theta}|X)}{\partial \theta_i^2}}}. \end{aligned} \tag{7}$$

The ML estimates and the standard errors of the object parameter estimates were used in the ASA.

Method

Simulation Study

We used R (Version 3.3.1) for this study (R Core Team, 2018). The R code for data simulation, both confirmatory and exploratory analyses, visualization of results, and deciding on the number of repetitions can be found in the

Supplementary Material in the online version of the article. The BTL model was used for both data simulation and data analysis.

First, we varied the selection algorithm, using both the ASA and a baseline algorithm to which the results of the newly developed algorithm were compared. In the baseline algorithm, which is the semi-random selection algorithm (SSA), a pair of objects is randomly selected with the constraint that the objects in the pair were not previously compared to each other. After the two selected objects were compared, the outcome was added to the data, and the selection algorithm was repeated until the predefined stopping criterion was reached. As the final step, the object parameter estimates and the corresponding standard errors were computed.

Second, we varied the number of objects N. We used N equal to 20, 25, 30, and 100 objects. These numbers represent three possible numbers of students in a class and one possible number of students in the same year of a school. We focused on these (small) sample sizes, which correspond with applications of pairwise comparison set up at a class level or a school level. Obviously, larger scale applications are also possible.

Third, we varied the number of comparisons performed by means of the proportion of the total number of unique comparisons. The total number of unique comparisons equals $N(N-1)/2$. We varied the proportion of the total number of unique comparisons that were used, denoted $C = 0.1\ (0.1)\ 1$. The condition $C = 1$ corresponds with a full design and can therefore be used as a benchmark.

For each of the $2 \times 4 \times 10$ (Algorithm $\times$ Number of Objects $\times$ Proportion of Comparisons) $= 80$ design cells, we drew N object parameters from the standard normal distribution. We used the conventional standard normal distribution because the ASA can be applied in a wide variety of contexts, and a previous article that reported unbiased distributional properties, resulting from non-adaptive pairwise comparison, reported different standard deviations for different samples (Van Daal et al., 2017), indicating that various standard deviations may be plausible. Because the object parameter estimates have a mean of 0 as a constraint for model identification, we rescaled the object parameters to have a sample mean of 0 as well. Subsequently, the probabilities that the objects are rated higher on the latent variable scale than other objects were computed by inserting their true (simulated) parameters in the BTL model. For example, for the standard normal distribution, an object with a simulated attribute value 1 *SD* above the mean of all objects will be preferred to an object with a simulated attribute value at the mean with a probability of $\exp(1-0)/[1-\exp(1-0)] = .73$.

In each cell, for each comparison, two objects were selected based on the selection algorithm. The comparison of the two objects was simulated by comparing a uniform random value between 0 and 1 to the probability that object i is

preferred to object *j*. Object *i* was chosen if the random value was smaller than the probability value, and object *j* was chosen otherwise. In the conditions involving the ASA, the object parameter estimates and the corresponding standard errors were computed. These steps were repeated until the maximum number of comparisons in the cell was reached. After reaching the maximum number of comparisons, object parameter estimates and standard errors were computed. Lastly, we computed the parameter uncertainty, the accuracy of ordering, and the reliability of the scale. This procedure, starting by drawing N object parameters for all cells, was repeated 400 times per cell. To determine the number of repetitions, we did a small simulation study for the cells with the highest variability of two evaluation criteria, benchmark reliability and Spearman's rank coefficient. These cells were the combinations of conditions $N = \{20, 25, 30\}$, $C = 0.1$, and algorithm {ASA and SSA}. The number of repetitions for which the standard errors of the benchmark reliability and Spearman's rank coefficient were below .01 for these cells was 384, which was rounded to 400 repetitions for the entire simulation study.

Evaluation Criteria

Uncertainty of parameters. We evaluated the uncertainty of the parameters using the standard error of the object parameter estimates. We expected that the standard errors were smaller for larger proportions of comparisons, and because the number of unique comparisons grows multiplicatively with the number of objects, expressed by the formula $N(N-1)/2$, we expected this effect to be larger for larger numbers of objects. In addition, we expected that the standard errors were larger using the SSA than the ASA. This difference was expected to be more pronounced for objects at the ends of the latent variable scale and less pronounced in the middle of the scale because objects at the ends of the scale usually have larger standard errors and therefore show larger possible gains.

Accuracy of ordering. The object order based on the object parameter estimates was compared to the object order in the generating model using Spearman's rank coefficient ρ, which is equal to Pearson's product–moment correlation between the estimated rank order of the objects and the object rank order in the generating model.

Reliability. We used two measures of reliability. First, we used the squared correlation between the object parameters used in the generating model and the object parameter estimates based on the data, which we refer to as the benchmark reliability. Let θ be the object parameter in the generating model and let $\hat{\theta}$ be the object parameter estimate. The benchmark reliability can then be computed as

$$\rho_{\hat{\theta}\hat{\theta}'} = \mathrm{cor}\left(\theta, \hat{\theta}\right)^2.$$

Second, we used the commonly used SSR estimate. Let $S^2(\theta)$ be the estimated true variance of the object parameters in the generating model, and let $\text{MSE}\left[SE\left(\hat{\theta}\right)\right]$ be the mean of the squared standard errors corresponding to the object parameter estimates. The SSR is computed as follows:

$$\text{SSR} = \frac{S^2(\theta)}{S^2\left(\hat{\theta}\right)},$$

where

$$S^2(\theta) = S^2\left(\hat{\theta}\right) - \text{MSE}\left[SE\left(\hat{\theta}\right)\right],$$

that is, the observed variance minus an error term (Bramley, 2015).

An increasing proportion of comparisons was expected to increase reliability by decreasing the standard errors of the object parameter estimates. We also expected the reliability to be higher using the ASA rather than using the SSA due to smaller standard errors of the parameter estimates. We expected this difference to be highest for a proportion of comparisons of 0.5 because the two algorithms have most selection degrees of freedom for this proportion of comparisons, at the compromise between the number of comparisons (i.e., opportunities of selection in performed comparisons), and the restriction that the comparisons must be unique (i.e., opportunities of selection in comparisons to be performed).

Results

Confirmatory Analyses

For $N = 30$, Figure 1 displays means of standard errors (dots) for ML estimates within ranges of 0.5 units of the latent variable scale. The figure shows that, except for proportions of comparisons equal to 0.1 and 1, the standard errors of the object parameter estimates at both ends of the latent variable scale were smaller when the ASA was used than when the SSA was used. Across the panels, it can be seen that the differences of the standard errors between the SSA and ASA decreased for increasing proportions of comparisons. This trend can be explained by the restriction that all unique comparisons could be made only once, which resulted in the algorithms having fewer degrees of freedom to select comparisons as C increased. Compared to the full design ($C = 1$), the SSA showed larger differences between standard errors found in the middle and the ends of the scale for lower proportions of comparisons, whereas the ASA showed smaller differences in standard errors. In general, for $C < 0.5$, the loss of precision compared to the full design was large. The ASA produced higher standard errors than the SSA in the middle of the scale, but the difference was negligible compared to the reduction of standard errors at both ends of the scale. The

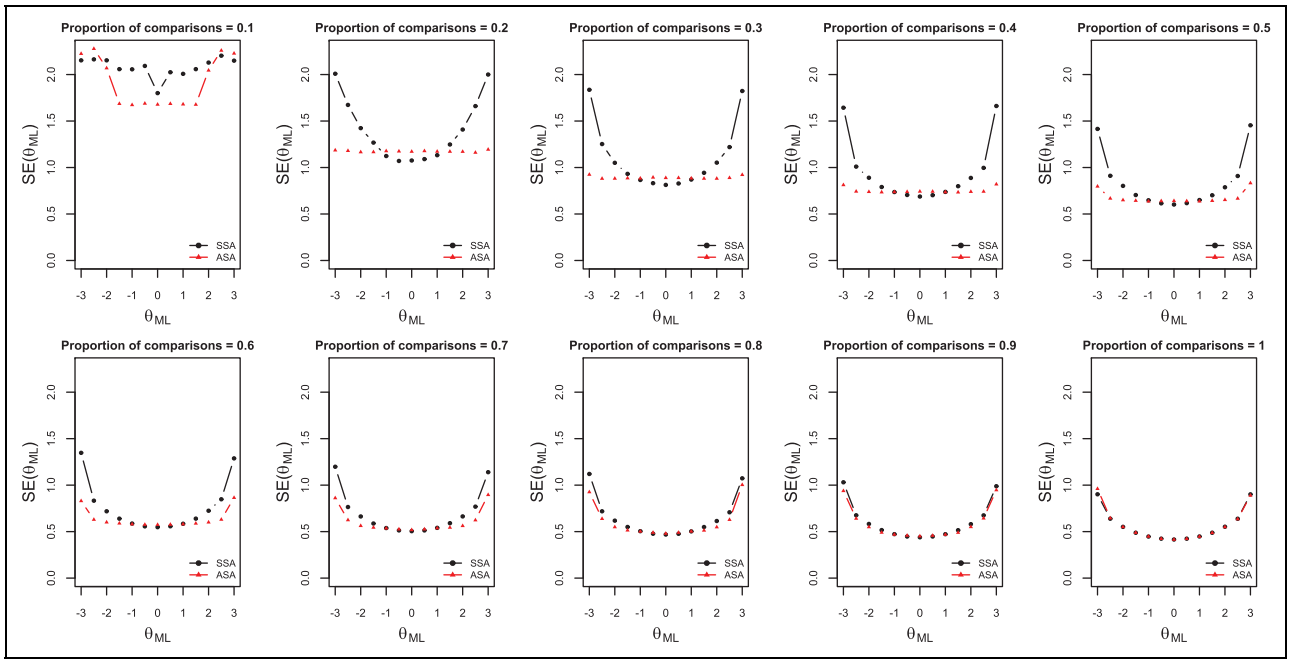

FIGURE 1. *Standard errors of object parameter estimates for both selection algorithms, for* $\mathrm{N} = 30$*, and for different proportions of comparisons. Means of standard errors (dots) for maximum likelihood estimates within ranges of .5 units of the latent variable scale.*

columns of Figure 2 show that for larger numbers of objects, the difference between the algorithms is smaller in the middle of the scale and larger at both ends of the scale, and this result is displayed for $C = 0.3$ and $C = 0.8$. However, the overall pattern of the differences between the algorithms was similar for all numbers of objects.

In Figure 1, the standard errors for $C = 0.1$ (upper left panel) showed a downward spike in the middle of the latent variable scale for the SSA, which was caused by some objects having perfect or zero scores. The nonaggregated estimates in Figure 3 illustrate this underlying cause. The dots above the "gap" in the left-hand panel of Figure 3 show that the objects with perfect or zero scores had parameter estimates above or below zero, respectively, but also that they had large standard errors. The lower (or higher) the parameter estimates of the objects to which they were preferred (or not preferred), the closer the parameter estimates of the objects with perfect (or zero) scores were to zero, and the higher their standard errors were. The downward spike in Figure 1 thus occurred due to the absence of objects with perfect or zero scores at latent variable estimates of zero. The right-hand panel of Figure 3 shows that the ASA did not suffer from perfect or zero scores this badly.

Both panels of Figure 4 show that Spearman's rank correlation was higher as the proportion of comparisons was larger. Similarly, the reduction in rank correlation compared to the full design ($\rho_{C=1} = .85$ for $N = 20$ and $\rho_{C=1} = .97$ for $N = 100$) was lower as the proportion of comparisons was larger, with small differences for $C > 0.6$ when $N = 20$ and for $C > 0.3$ when $N = 100$. Given the proportion of comparisons, the rank correlation was higher for larger sample sizes, which is reflected in the difference between the panels. The explanation is that the number of unique comparisons is higher for larger sample sizes, and therefore, the same proportion represents more comparisons for larger sample sizes. The narrower confidence intervals for larger sample sizes can be explained in the same manner. An unexpected result was the absence of a difference of Spearman's ρ between the two algorithms, indicated by the closeness of the black and red lines, meaning that the ASA does not produce higher rank order accuracy than the SSA.

For benchmark reliability, the same trends as for the rank order accuracy were found. Figure 5 shows that the benchmark reliability increased when proportions of comparisons increased, and the figure shows between panels that benchmark reliability also increased when sample size increased. Compared to the full design ($\text{Reliability}_{C=1} = .78$ for $N = 20$ and $\text{Reliability}_{C=1} = .95$ for $N = 100$) large differences occurred for $C < 0.8$ when $N = 20$ and for $C < 0.5$ when $N = 100$. ASA and SSA produced similar benchmark reliabilities. However, the SSR showed a difference between the two algorithms. The dashed lines of Figure 5 show that using the SSA, the SSR on average underestimated benchmark reliability, but the SSR overestimated reliability for most proportions of comparisons when the ASA was used. This overestimation is undesirable for a

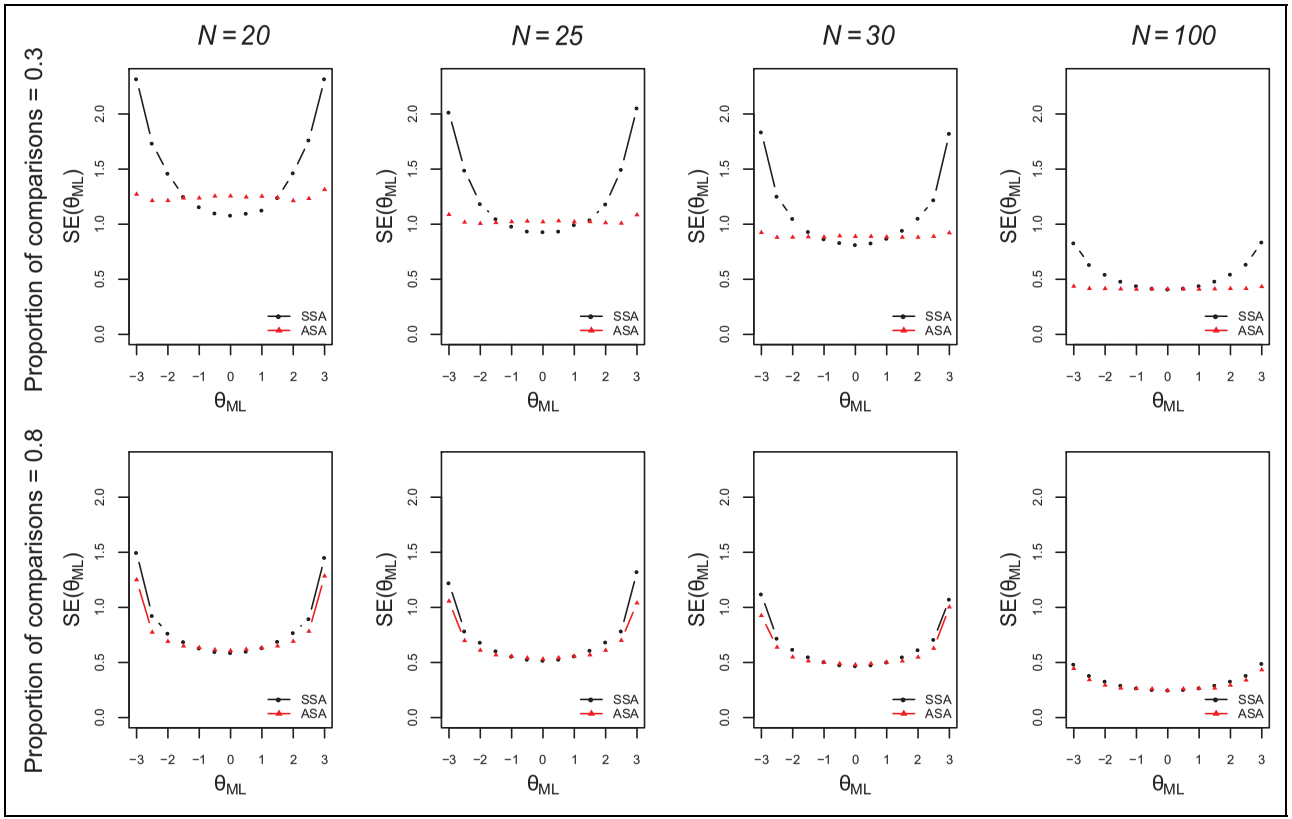

FIGURE 2. *Standard errors of object parameter estimates for both selection algorithms, for different* N, *and for proportions of comparisons of .3 and .8. Means of standard errors (dots) for maximum likelihood estimates within ranges of .5 units on the latent variable scale.*

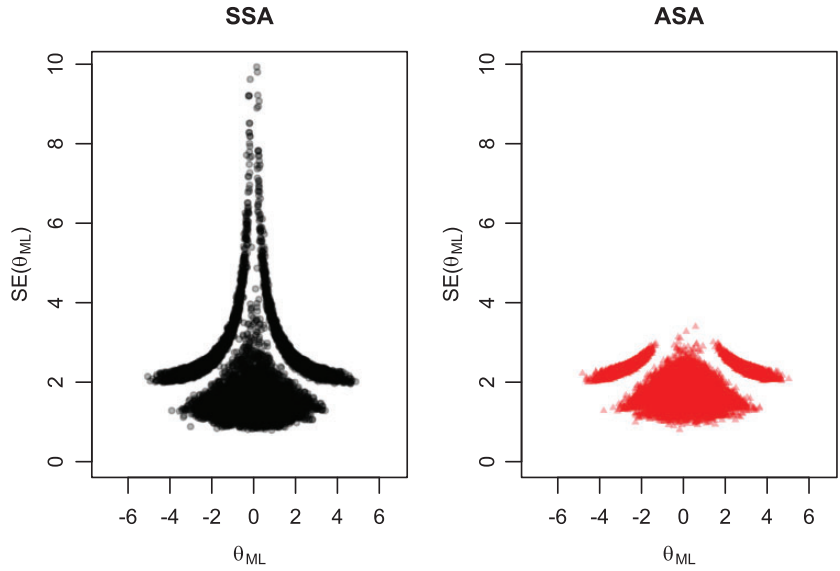

FIGURE 3. *Standard errors of object parameter estimates for both selection algorithms, for* N = *30, and for* C = *0.1.*

reliability estimate because reliability estimates must not suggest that the measurement quality is higher than it actually is (Sijtsma, 2009).

The black and red lines in Figure 6 show that an average of 20 to 22 comparisons per object are required to obtain a reliability of .80. The gray and pink lines show that more than 30 comparisons per object are required to obtain a lower bound of the 90% confidence interval of at least .80. Figure 6 indicates that the proportion of total unique comparisons shows a trend, but it seems that the mean number of comparisons per object shows a clearer relation with reliability. This result is especially interesting because this result can be directly applied to large-scale assessment. This is not the case when looking at the proportion of the total number of possible comparisons, since this statistic depends on the number of objects.

Exploratory Analyses

We conducted exploratory analyses to gain further understanding of several results of the confirmatory analyses. In our confirmatory analyses, we noticed that the standard errors of the object parameter estimates and the SSR were affected by the constraint that all comparisons must be unique when the proportion of comparisons was close to 1 or equal to 1. We investigated what the results would be if we would release the restriction that all comparisons must be unique,

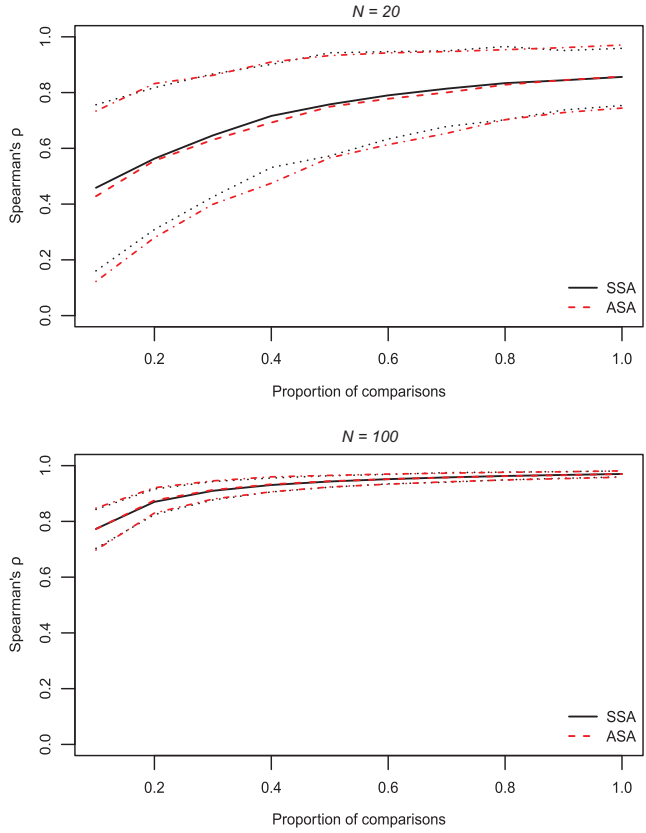

FIGURE 4. *Spearman rank correlation between true and estimated object rank order and 90% confidence interval for different proportions of comparisons.*

which corresponds with a situation involving multiple raters that agree perfectly. More specifically, the results are as if multiple raters performed independent comparisons using the same decision rule. We investigated the results of both the SSA and the ASA without this restriction in the following conditions: $N = \{20, 25, 30\}$ and $C = \{0.1\ (0.1)\ 1, 2\}$. It may be noted that, because the only restriction on randomness is removed, the unrestricted SSA is actually a fully random selection algorithm instead of a semi-random selection algorithm.

The standard errors of the object parameter estimates were smaller for the adaptive algorithm than the random algorithm for all proportions of comparisons (Figure 7). This effect was smaller for larger proportions of comparisons in the original simulation study. The SSR overestimated the benchmark reliability for all proportions of comparisons (Figure 8). In the original simulation study, the

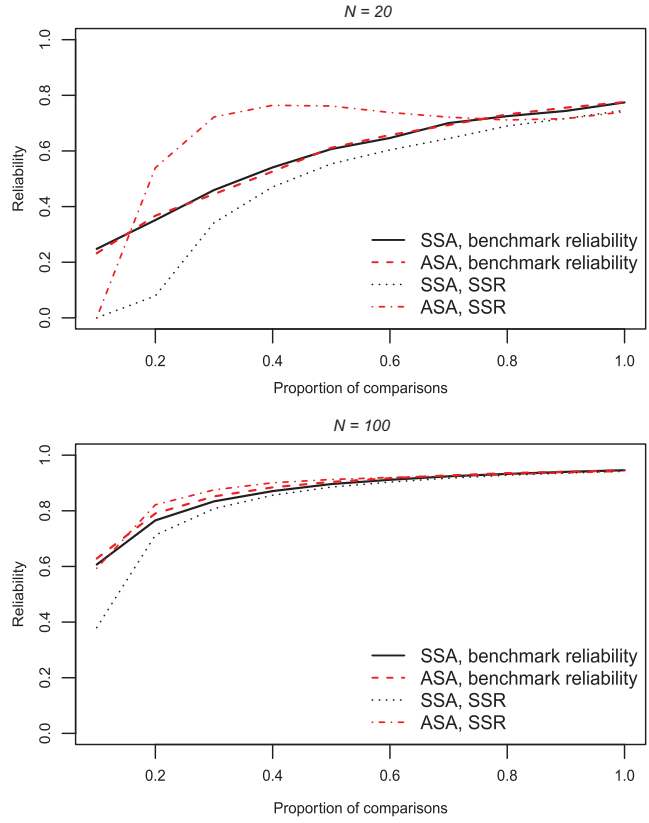

FIGURE 5. *Benchmark reliability and estimated reliability for different proportions of comparisons.*

SSR did not overestimate the reliability for large proportions of comparisons, which can be attributed to the unique comparison restriction.

Another result from the simulation study was that the SSR overestimated reliability when the ASA was used. Further inspection of the results showed that the variance of the object parameters was overestimated. This result was found for both algorithms, but overestimation was extremely large for the adaptive algorithm when C was small. To further investigate the mechanism producing this result, we fixed the object parameters in the generating model and tested the unrestricted SSA and ASA in the following conditions: $N = \{10, 20, 30\}$ combined with $C = 5$ and three different sets of object parameters in the generating model for condition $N = 3$ combined with $C = 20$. For these conditions, we used eight repetitions.

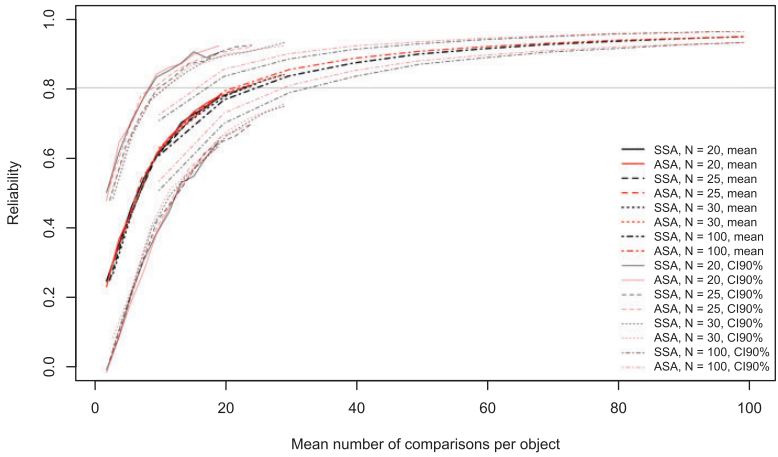

FIGURE 6. *Benchmark reliability and 90% confidence interval for different numbers of comparisons per object for different sample sizes.*

Figure 9 shows that for the first set of conditions ($C = 5$), the flat lines starting from 10 to 15 comparisons per object suggest that the estimated true variance converged after about 10 to 15 comparisons per object. However, for the adaptive algorithm, the horizontal lines above the value 1 suggest that the estimate occasionally converged to an incorrect variance estimate. This result shows that for the ASA, the overestimation of the variance is not by definition resolved asymptotically, which might also be a problem for other adaptive algorithms that overestimate the variance. For the second set of conditions ($N = 3$), we noticed that the object parameters of the three generating models produced different results. When the object parameters were close to each other, the location of these objects was estimated quite precisely, but the order of the objects and the variance of their parameters were not. The opposite results were found for both the SSA and the ASA when the object parameters were distant.

Discussion

The newly developed ASA produced smaller standard errors than the SSA. This result was found both for the version of the adaptive algorithm restricted by unique comparisons and for the unrestricted version. For ASA and SSA, the Spearman rank correlation and the benchmark reliability were similar, for both the restricted and unrestricted versions. On average, 20 comparisons per object are required for a benchmark reliability of at least .80. The SSR coefficient on average underestimated reliability when the SSA was used, but overestimated reliability when the ASA was used, and the overestimation grew larger when the ASA was unrestricted. A possible explanation is that using the ASA, the variance

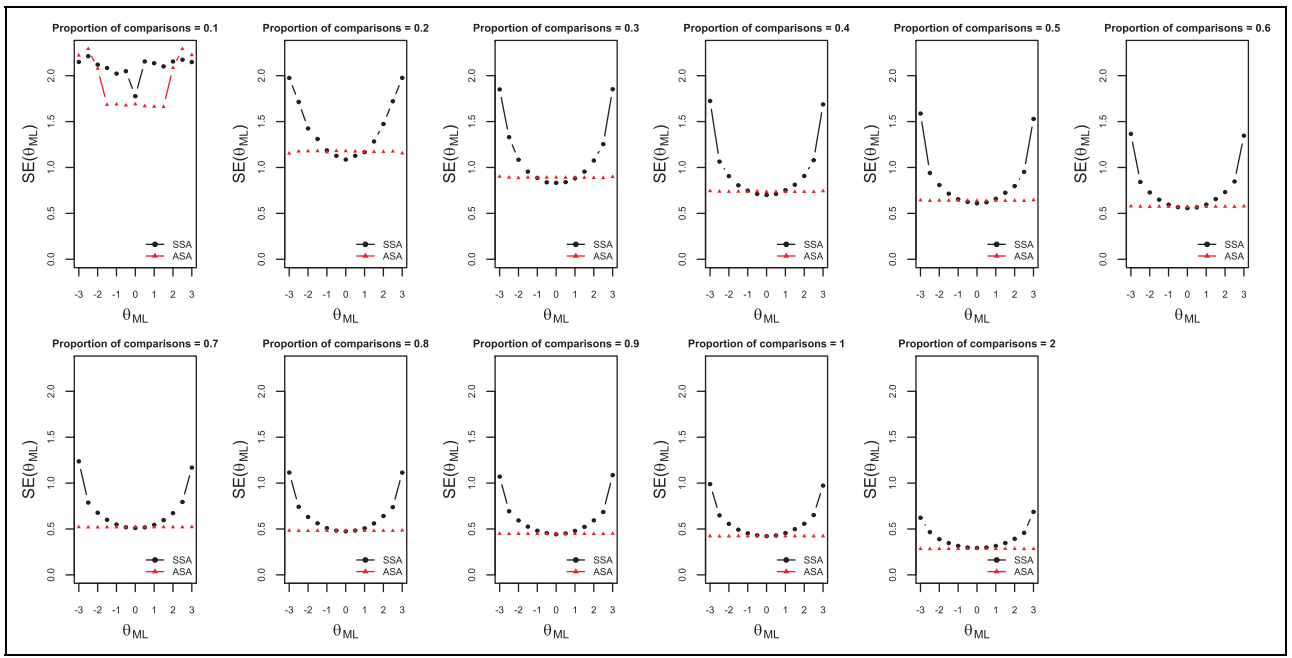

FIGURE 7. *Standard errors of object parameter estimates for both selection algorithms, not restricted to unique comparisons, for* N = *30 and for different proportions of comparisons. Means of standard errors (dots) for maximum likelihood estimates within ranges of .5 units on the latent variable scale.*

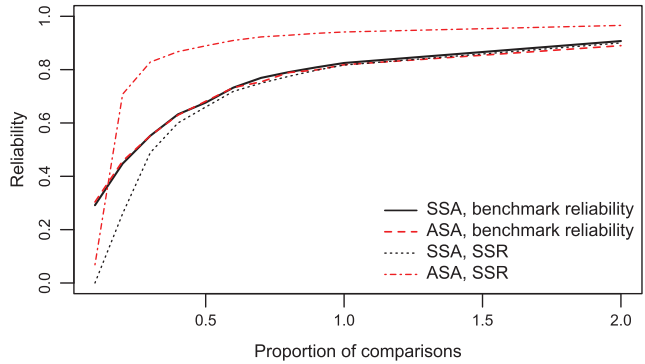

FIGURE 8. *Benchmark reliability and estimated reliability for varying values of* C *and* N = *30.*

of the object parameters was overestimated. These results support the suggestion of Bramley and Vitello (2018) that using an adaptive algorithm can lead to a spuriously inflated standard deviation of the object parameters, but the standard errors of the parameters can be genuinely reduced. Therefore, this conclusion probably applies to other adaptive pairwise comparison algorithms that lead to an inflated SSR coefficient as well.

When the object parameters in the generating model were close to each other, the location of these objects was estimated quite precisely, but the order of the objects and the variance of their parameters were not, while the opposite results were found when the object parameters in the generating model were distant. This result was found both for the SSA and the ASA and might also hold for other pairwise comparison algorithms. This conclusion may seem obvious but should be kept in mind when interpreting location parameter estimates or rank order estimates from a single sample.

This study contributes to adaptive pairwise comparison by proposing an ASA that takes the uncertainty of the parameters into account. The ASA can be used to decrease the standard errors of the object parameter estimates, hence to increase precision of object locations on the latent variable scale. The improvement holds for the entire group of objects but is largest for objects at both ends of the attribute scale. Therefore, one could argue that the improvement may have limited impact in practical situations. The ASA provides little to no advantage compared to the SSA with respect to reliability and rank order accuracy. Further research could develop an algorithm that focuses on increasing the reliability because in most situations, teachers may be interested in the rank order of students on an attribute scale rather than the location on the scale. In these situations, the focus lies on reliability instead of precision of parameter estimates. For example, a teacher may want to form groups of students for a group

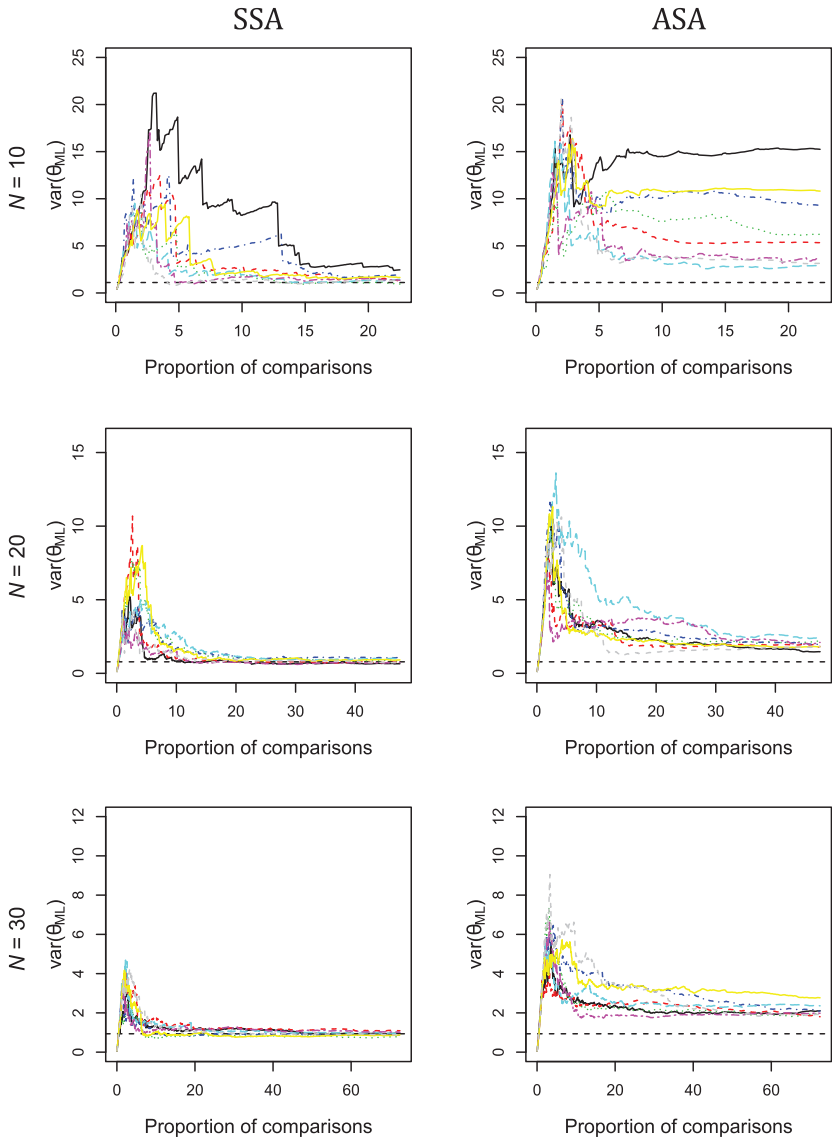

FIGURE 9. *Variance of object parameters for varying numbers of comparisons per object for eight replications.*

assignment based on their relative position in the class on this attribute. Forming groups may then be accomplished by grouping students ranked close together or grouping higher ranked students with lower ranked students.

Whereas previous studies used real data or simulated data without replications (Bramley, 2015; Bramley & Vitello, 2018; Pollitt, 2012), we used simulated data with 400 replications in various conditions. The simulated data allowed us to compare the SSR reliability coefficient with the benchmark reliability, and we found that adaptivity can lead to an inflated SSR coefficient. The large number of replications ruled out that sampling fluctuations explain the results, which was possible in previous research designs (Bramley, 2015).

This study focused on a design with a single rater that performed all comparisons, and the study did not investigate the influence of various raters on the performance of the algorithms. For high-stakes assessment, one rater would be undesirable. First, the burden on this rater would be high. Second, the subjectivity of the rater cannot be counterbalanced by the judgments of other raters. However, having one rater may not be a problem in a classroom situation with low-stakes assessment when the teacher is evaluating whether students understand what he or she has taught or when the evaluation is used to facilitate learning. Hence, our results might be valuable for these low-stakes situations.

Varying numbers of raters and percentages of rater agreement might be valuable when studying the algorithms for use in high-stakes assessment, but their inclusion would render the study design large and time-consuming. It would also require additional research in the different ways rater variance should be modeled. Therefore, we chose to illustrate how the adaptive algorithm technically performs using a single rater as a proof of concept and to illustrate which issues may arise when using this or a similar adaptive algorithm. Even though the influence of raters itself was not investigated, the effects of the algorithm, the number of objects, and the proportion of comparisons on the evaluation criteria can be generalized to the setting of multiple raters. This study can be used in future research as a baseline to investigate the influence of the number of raters in combination with rater agreement. For example, different degrees of rater agreement may be achieved by varying the preference probabilities of the objects for different raters, where larger differences between raters might increase parameter uncertainty and decrease rank order accuracy as well as both types of reliability for all algorithms.

The scale that is obtained from the pairwise comparisons could be used to test whether two objects significantly differ from each other on the attribute of interest. The standard errors of the object parameter estimates could be used to create confidence intervals around the object parameter estimates, which can in turn be used to test whether an object is different from another object. The smaller standard errors the adaptive algorithm produced would lead to smaller confidence intervals, which in turn would lead to higher statistical power. Unfortunately, in several conditions, the adaptive algorithm also led to an overestimated variance of the object parameters, so the differences between objects might be overestimated. Although the statistical power is higher with the adaptive algorithm, this overestimation may cause the power to be

overestimated as well, suggesting that the power is even higher when it is not. For this reason, and because the adaptive algorithm was not developed for this specific purpose, we do not advice significance testing for differences between the objects.

To conclude, for the same number of comparisons, the ASA developed in this study can be used to obtain estimates of objects on a latent variable that are more precise than when a random algorithm is used. However, because the SSR may overestimate reliability, one should be cautious to interpret the SSR coefficient when using adaptive pairwise comparisons. On average, about 20 comparisons per object are required for a reliability of .80, whether one uses an adaptive algorithm for pairwise comparison or not.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article

References

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2018). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design in Education*, *2018*, 1–23. doi:10.1007/s10798-018-9442-7

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: 1. The method of paired comparisons. *Biometrika*, *39*, 324–345.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgement. Cambridge assessment research report*. Cambridge, England: Cambridge Assessment.

Bramley, T., & Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *2018*, 1–16. doi:10.1080/0969594X.2017.1418734.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, *27*, 412–433. doi:10.1214/12-STS396

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, *32*, 384–406. doi:10.1214/aos/1079120141

Jones, I., & Alcock, L. (2013). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*, 1774–1787. doi:10.1080/03075079.2013.821974

Laming, D. (2004). *Human judgment: The eye of the beholder*. London, England: Thomson Learning.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Mayer, S. (2017). Comparative judgement as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative practices for higher education assessment and measurement* (pp. 119–138). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-0531-0.ch007.

Luce, R. D. (1959). *Individual choice behaviours: A theoretical analysis*. New York, NY: Wiley.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice*, *21*, 205–220. doi:10.1080/0969594X.2013.868341

Pollitt, A. (2004). *Let's stop marking exams*. Presented at the IAEA Conference, Philadelphia, PA.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*, 281–300. doi:10.1080/0969594X.0962012.0665354.

Pollitt, A. (2015). *On "reliability" bias in ACJ: Valid simulation of adaptive comparative judgement. Cambridge Exam Research*, Cambridge, England.

Rangel-Smith, C., & Lynch, D. (2018). *Addressing the issue of bias in the measurement of reliability in the method of adaptive comparative judgment*. Paper presented at the 36th International Pupils' Attitudes Towards Technology Conference, Westmeath, Ireland. Retrieved from http://terg.ie/index.php/patt36-proceedings/

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 45–64). Nathan, Australia: Griffith Institute for Higher Education. doi:10.1007/978-1-4020-8905-3_4

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design in Education*, *22*, 205–226. doi:10.1007/s10798-011-9194-0

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120. doi:10.1007/s11336-008-9101-0.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, *29*, 211–223. doi:10.1080/08957347.2016.1171769

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, *2016*, 1–16. doi:10.1080/0969594X.2016.1253542.

Van Daal, T., Lesterhuis, M., Coertjens, L., Van de Kamp, M. T., Donche, V., & De Maeyer, S. (2017). The complexity of assessing student work using comparative judgment: The moderating role of decision accuracy. *Frontiers in Education*, *2*, 1–13. doi:10.3389/feduc.2017.00044

Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer-Verlag.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., … Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). New York, NY: Routledge.

Authors

ELISE A. V. CROMPVOETS is a PhD student at both the Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands and at Cito, Amsterdamseweg 13, 6814 CM Arnhem, The Netherlands; email: e.a.v.crompvoets@uvt.nl, elise.crompvoets@cito.nl. This research project is part of her dissertation about pairwise comparison for educational measurement.

ANTON A. BÉGUIN is a director of central tests and examinations at Cito, PO Box 1034, 6801 MG Arnhem, the Netherlands; email: anton.beguin@cito.nl. He received his doctorate from Twente University where he did research on multidimensional item response theory and test equating.

KLAAS SIJTSMA is a full professor at the Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands; email: k.sijtsma@uvt.nl. He received his PhD in psychology at the Rijksuniversiteit Groningen in 1988. His scientific interest concentrates on the measurement of individual differences with respect to psychological constructs.

Manuscript received August 9, 2018
First revision received December 24, 2018
Second revision received July 4, 2019
Accepted September 20, 2019